Video: Protein-Protein Interactions
Credit: Mauro Martino

Network Science

Cristian Candia-Castro Vallejos, Ph.D.

cristiancandia@udd.cl

Director Magister en Data Science UDD
Profesor Investigador, Facultad de Ingeniería, UDD
External Faculty Northwestern Institute on Complex Systems,
Kellogg School of Management, Northwestern University

Power Law

Estas diapositivas se basan en la presentación original del
Prof. Albert-László Barabási, de Northeastern University, con autorización.
El contenido ha sido traducido para su uso en este curso.

Gráficos de Leyes de Potencia

POWER-LAW DISTRIBUTIONS IN EMPIRICAL DATA

AARON CLAUSET*, COSMA ROHILLA SHALIZI†, AND M. E. J. NEWMAN‡

Abstract. Power-law distributions occur in many situations of scientific interest and have significant consequences for our understanding of natural and man-made phenomena. Unfortunately, the detection and characterization of power laws is complicated by the large fluctuations that occur in the tail of the distribution—the part of the distribution representing large but rare events—and by the difficulty of identifying the range over which power-law behavior holds. Commonly used methods for analyzing power-law data, such as least-squares fitting, can produce substantially inaccurate estimates of parameters for power-law distributions, and even in cases where such methods return accurate answers they are still unsatisfactory because they give no indication of whether the data obey a power law at all. Here we present a principled statistical framework for discerning and quantifying power-law behavior in empirical data. Our approach combines maximum-likelihood fitting methods with goodness-of-fit tests based on the Kolmogorov-Smirnov statistic and likelihood ratios. We evaluate the effectiveness of the approach with tests on synthetic data and give critical comparisons to previous approaches. We also apply the proposed methods to twenty-four real-world data sets from a range of different disciplines, each of which has been conjectured to follow a power-law distribution. In some cases we find these conjectures to be consistent with the data while in others the power law is ruled out.

Gráfico Power-law

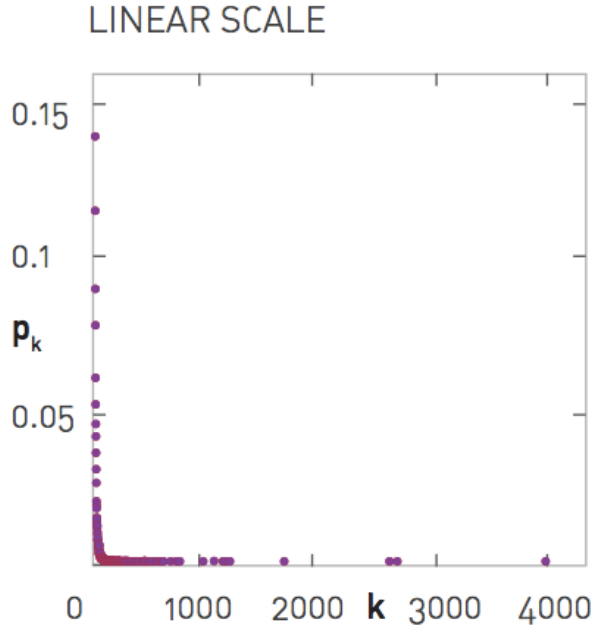
$$(k_1, k_2, \ldots, k_{N-1}, k_N)$$

1. Crea un histograma

$$p_k = \frac{N_k}{N}$$

Cuenta del numero de nodos que tienen grado 1, grado 2, etc.

- Poner leyenda a los ejes
- Números visibles

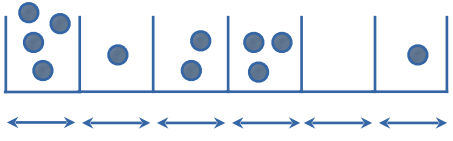


A.-L. Barabási. (2016). *Network Science, Chapter 4*.

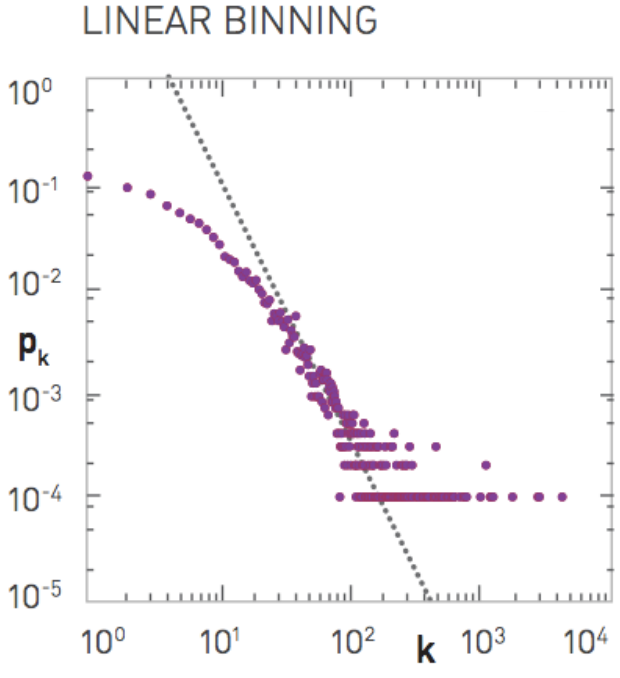
Gráfico Power-law

$$(k_1, k_2, \ldots, k_{N-1}, k_N)$$

2. Grafica en escala log-log



espacios equivalentes



A.-L. Barabási. (2016). *Network Science, Chapter 4*.

Gráfico Power-law

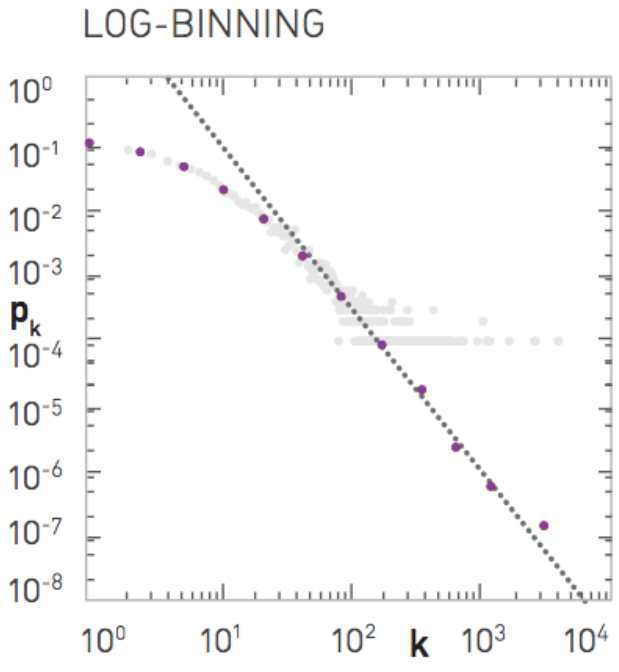
$$(k_1, k_2, \ldots, k_{N-1}, k_N)$$

3. Ajustes estéticos

Eliminar ruido para entender la tendencia de la cola

Opción 1:

Binning logarítmico



A.-L. Barabási. (2016). *Network Science, Chapter 4*.

Gráfico Power-law

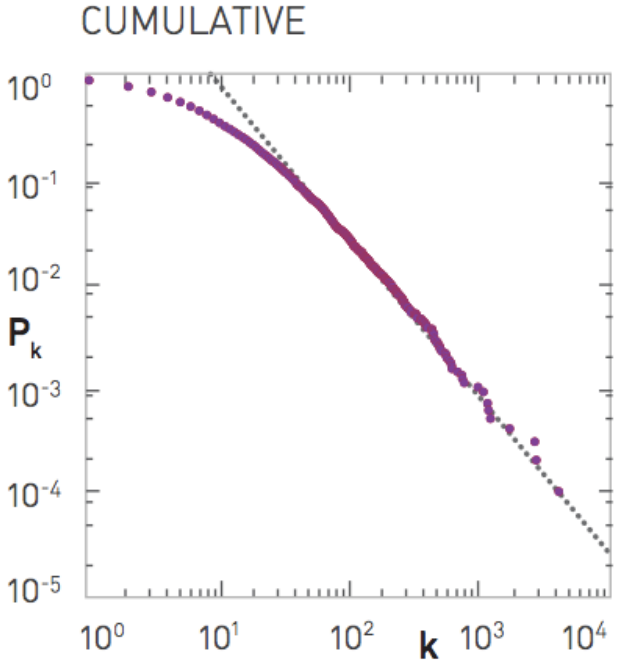
$$(k_1, k_2, \ldots, k_{N-1}, k_N)$$

3. Ajustes estéticos

Eliminar ruido para entender la tendencia de la cola

Opción 2:

Distribución acumulada



A.-L. Barabási. (2016). *Network Science, Chapter 4*.

Elegir el binning correcto

Example 6.6. (Choice of the optimal binning) Suppose the bottom of the lowest bin is at $k_{\min}$ and the ratio of the widths of successive bins is a. Then the n-th bin extends from $x_{n-1} = k_{\min}a^{n-1}$ to $x_n = k_{\min}a^n$, and the expected number of samples falling in this interval is:

$$\int_{x_{n-1}}^{x_n} p(k)dk = c\int_{x_{n-1}}^{x_n} k^{-\gamma}dk$$
$$= c\frac{a^{\gamma-1}-1}{\gamma-1}(k_{\min}a^n)^{1-\gamma}.$$

Thus, so long as $\gamma > 1$ which, as shown in Table 6.1, is true for most of the degree distribution of real networks, the number of samples per bin goes down as n increases as $a^{(1-\gamma)n}$. Consequently, the bins in the tail will have more statistical noise than those that precede them.

We can further reduce the fluctuations in the tails with a different binning in which the ratio of the width of successive bins is not necessarily constant. Such an *ideal binning* would indeed be obtained by imposing that the expected number of samples in each bin is *exactly* the same. Mathematically, this is equivalent to imposing that:

$$\int_{x_{n-1}}^{x_n} p(k)dk = \frac{1}{N_b}$$

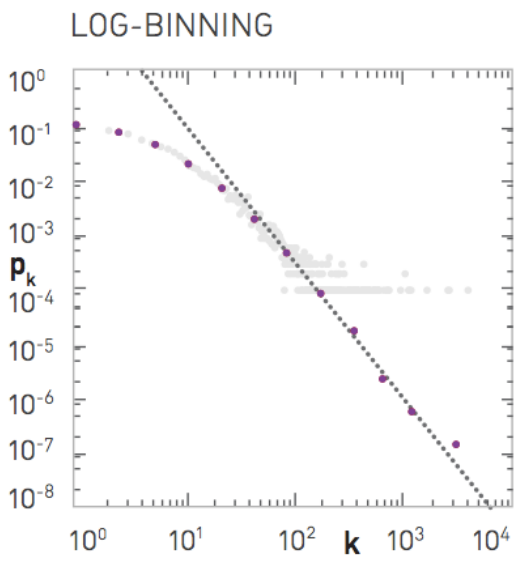
where N_b is the number of bins. We thus obtain $(\frac{x_{n-1}}{k_{\min}})^{1-\gamma} - (\frac{x_n}{k_{\min}})^{1-\gamma} = \frac{1}{N_b}$ which gives:

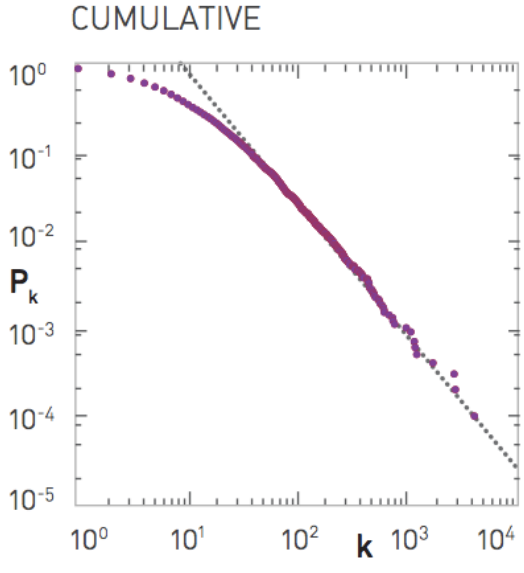
$$x_n = k_{\min}(1 - \frac{n}{N_b})^{-\frac{1}{\gamma-1}} \qquad (6.24)$$

This choice is the one that minimizes the overall fluctuations for the same number of bins and samples. However, the problem with this choice is that the two sides of a bin are in general not integer numbers. □

V. Latora. (2014)

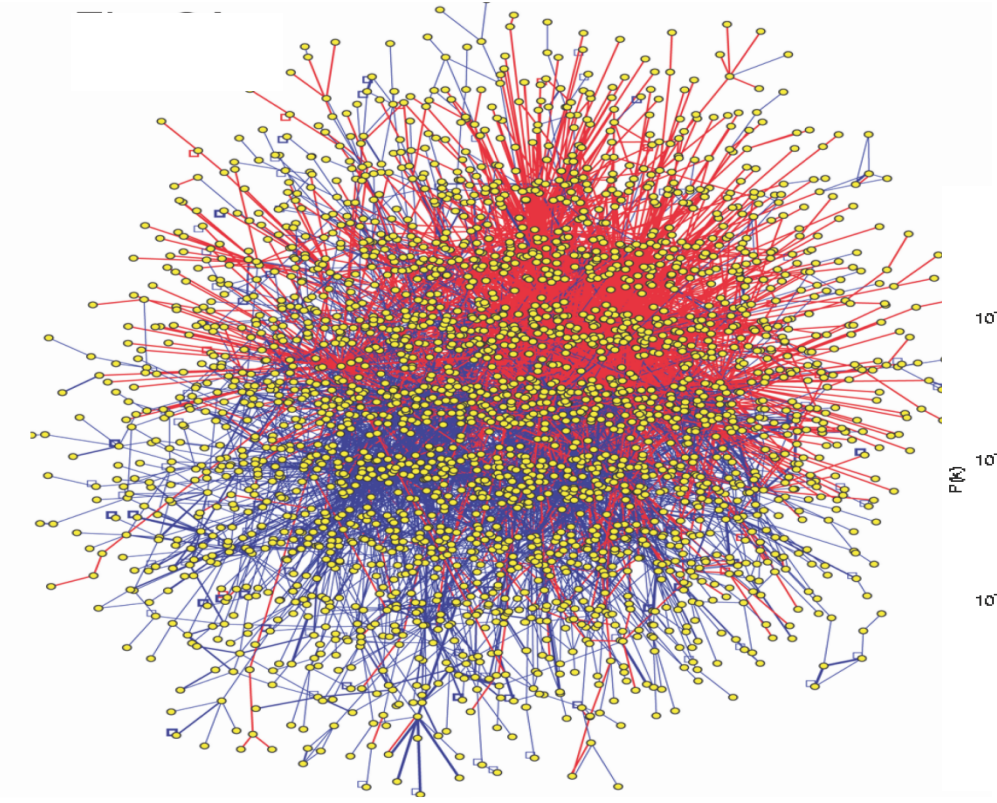
Elegir el binning correcto





A.-L. Barabási. (2016). *Network Science, Chapter 4*.

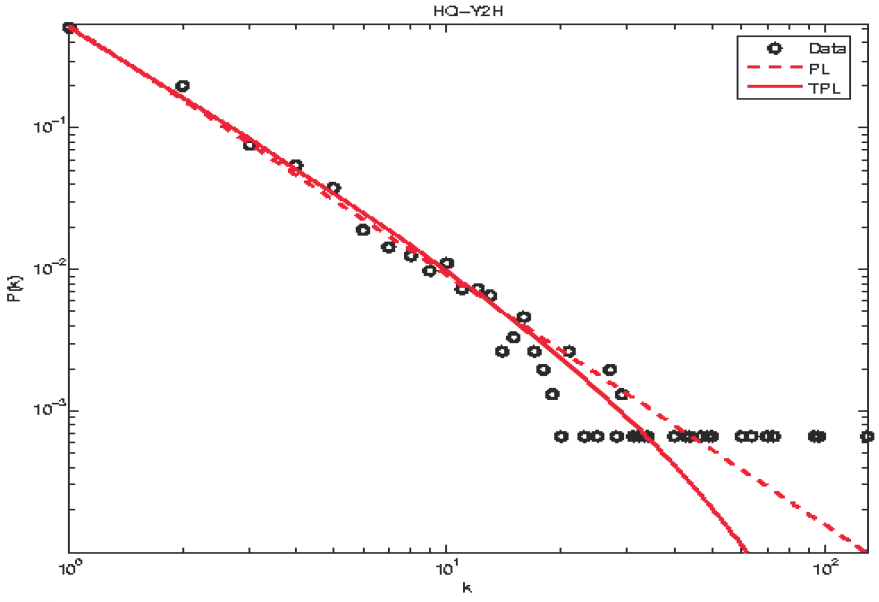
RED DE INTERACCION HUMANA



2,800 Y2H interactions

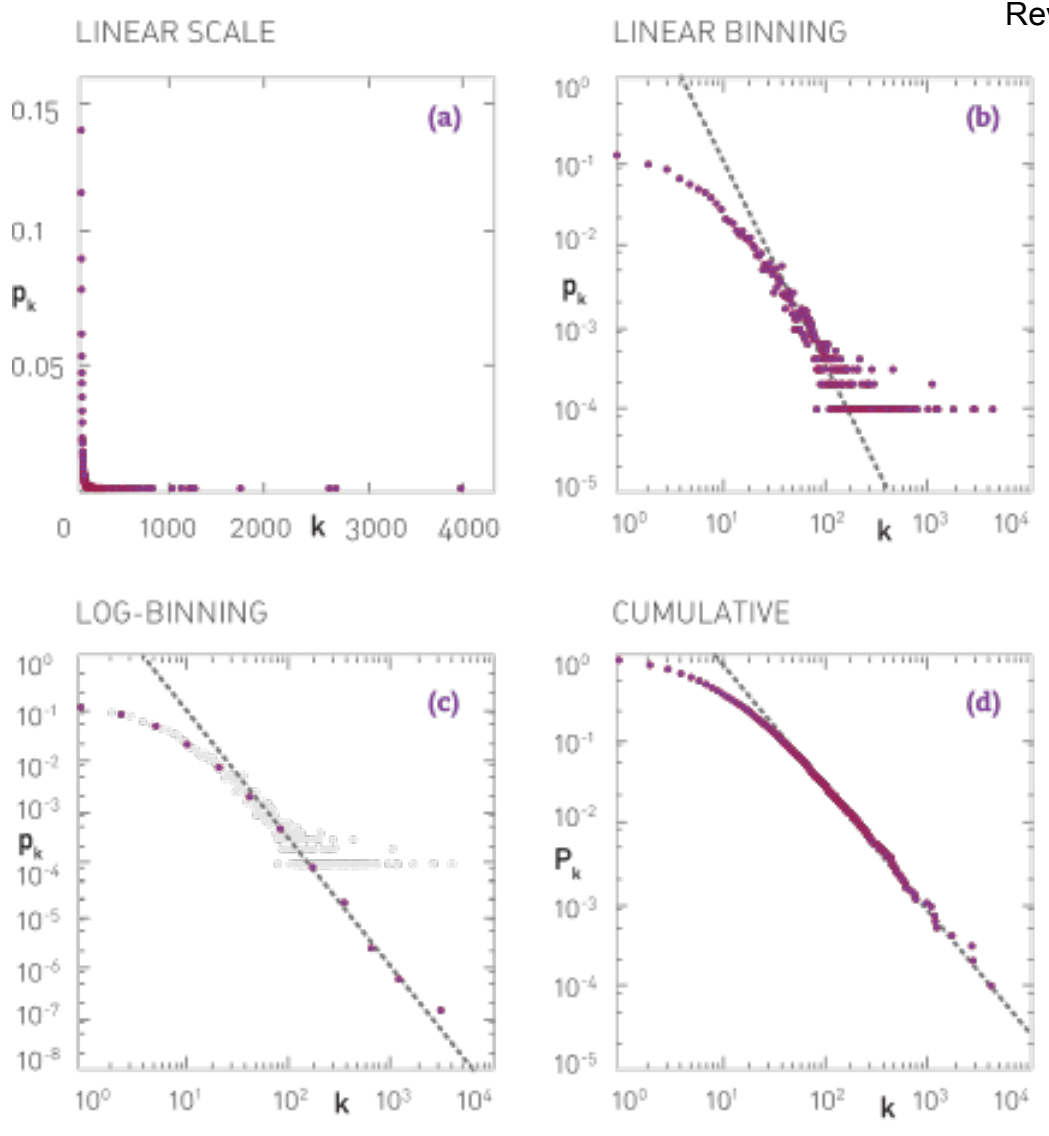
4,100 binary LC interactions

(HPRD, MINT, BIND, DIP, MIPS)



Rual et al. Nature 2005; Stelze et al. Cell 2005

Revisar: http://networksciencebook.com/chapter/4#advanced-b



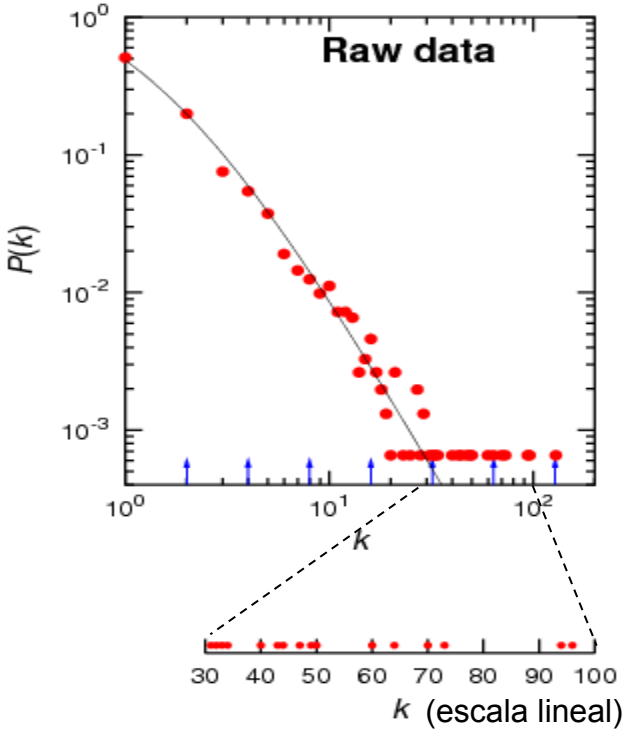
Usa gráficos log-log

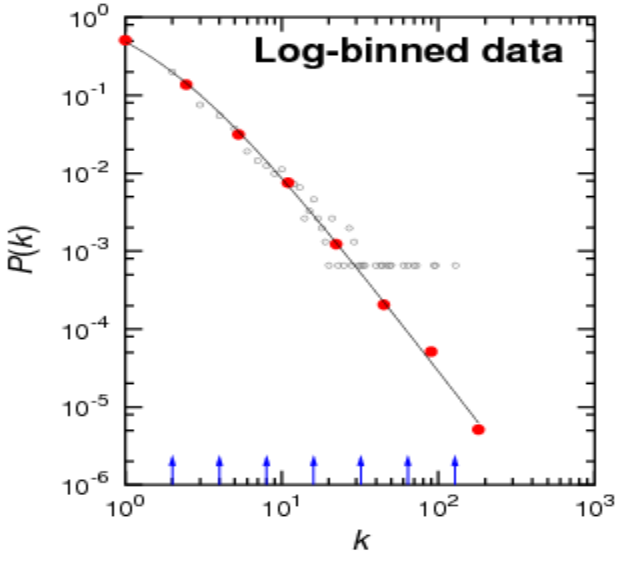
Evita el bining lineal

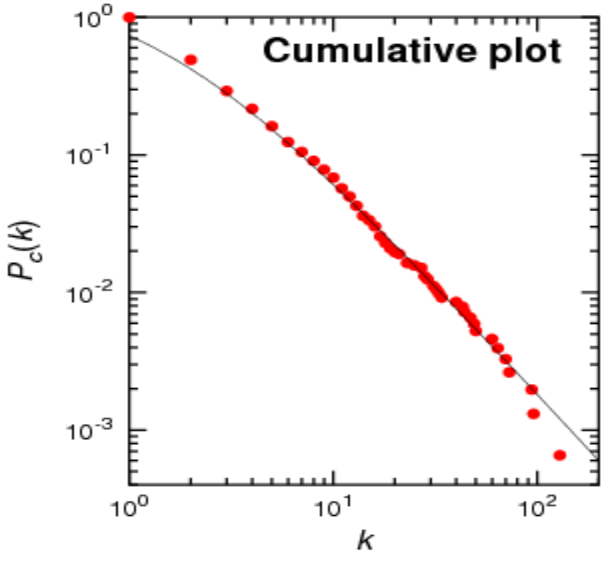
Usa bining logarítmico

Usa la distribución acumulada

DATA DE INTERACCION HUMANA POR RUAL ET AL.







$$P(k) \sim (k+k_0)^{-\gamma}$$

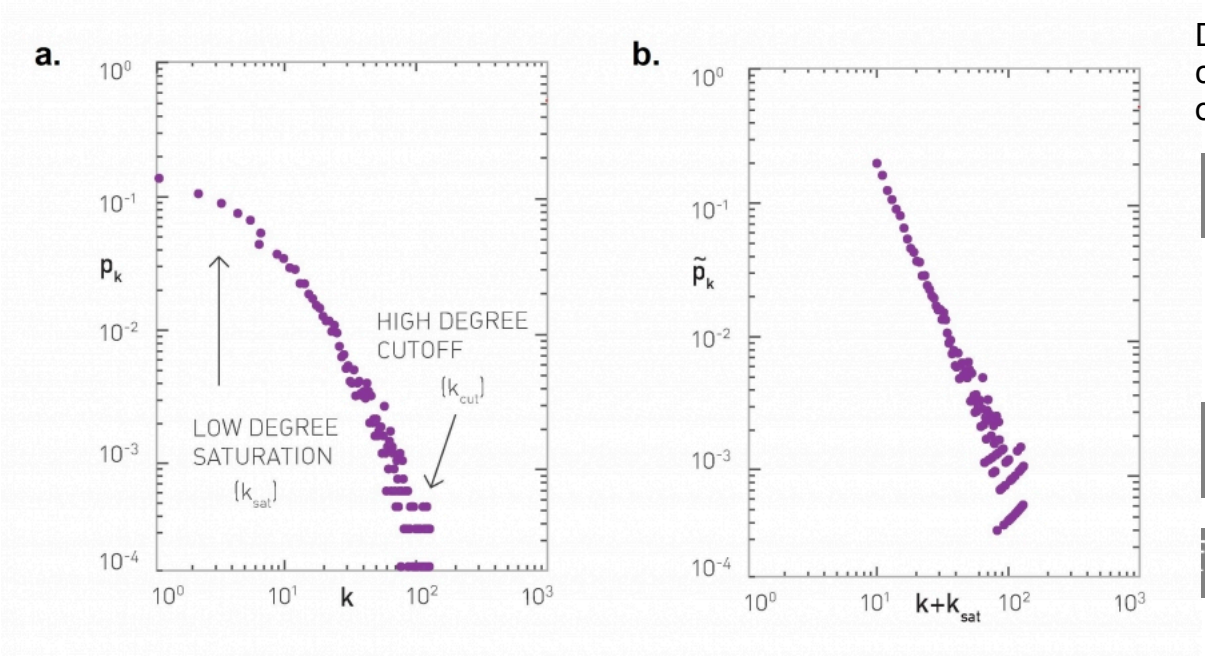
$$k_0 = 1.4,\ \gamma = 2.6.$$

Fallas comunes

Reescalando la distribución del grado

En redes reales, la distribución de grados con frecuencia se desvía de una ley de potencia pura al mostrar una saturación para grados bajos y un corte en grados altos.

Al trazar la función p reescalada en ($k + k_{sat}$), la distribución de grados sigue una ley de potencia para todos los grados.



Dada la presencia generalizada de tales cortes, la distribución de grados se ajusta ocasionalmente a:

$$p_x = a(k + k_{sat})^{-\gamma} \exp\left(-\frac{k}{k_{cut}}\right) \qquad (4.39)$$

Para extraer la extensión completa de la escala se grafica:

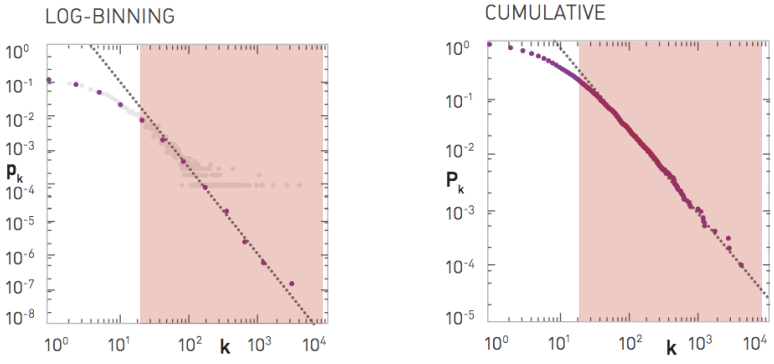
$$\tilde{p}_x = p_x \exp\left(\frac{k}{k_{cut}}\right) \qquad (4.40)$$

$\tilde{k} = k + k_{sat}$.

$\tilde{p} \sim \tilde{k}^{-\gamma}$

Revisar: http://networksciencebook.com/chapter/4#advanced-c

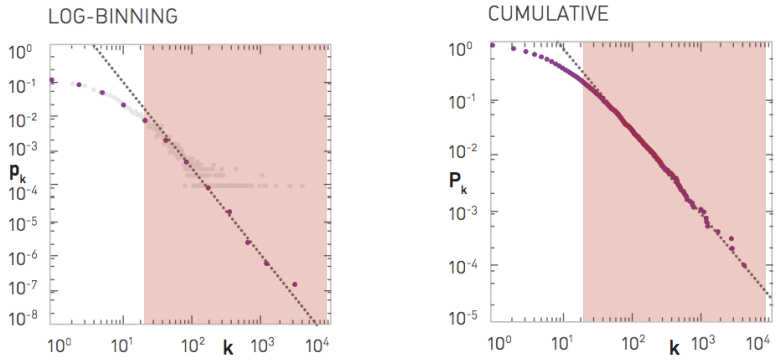
Ajuste de Power-law



Ajustar con el meted de los cuadrados minimus en la portion del grafico que parece mas lineal en la escala log-log

A.-L. Barabási. (2016). *Network Science, Chapter 4*.

Ajuste de Power-law



La pendiente del ajuste proporciona el exponente (log-binning)

$$\log p_k = -\gamma \log k + \text{constant}$$

La pendiente del ajuste proporciona el exponente +1 (cumulative)

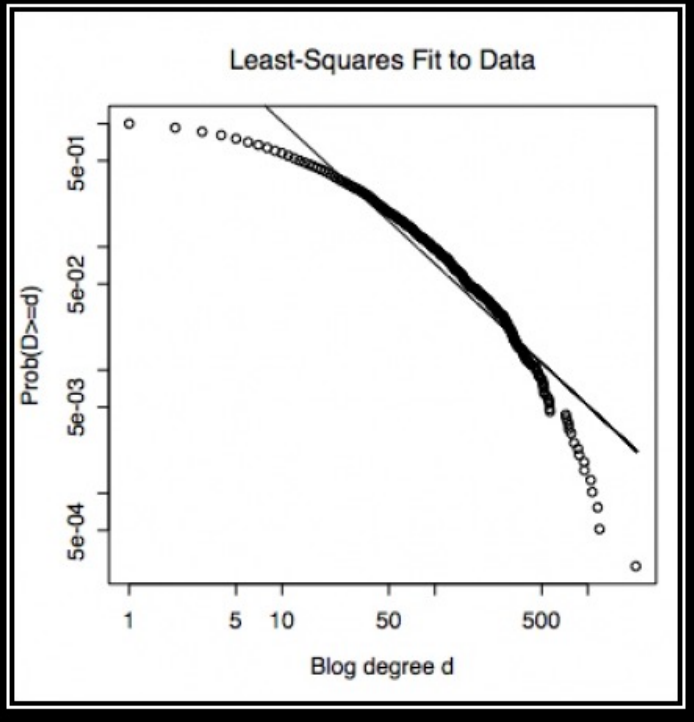
$$\log P_k = (-\gamma + 1) \log k + \text{constant}$$

A.-L. Barabási. (2016). *Network Science, Chapter 4*.

Ajuste de Power-law

Por qué este ajuste es incorrecto?

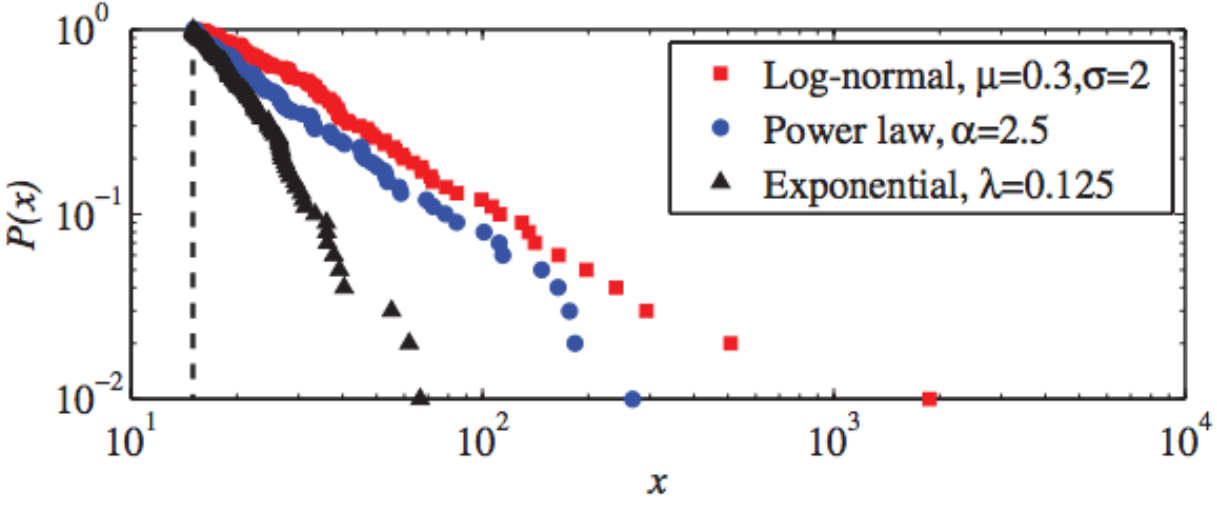
- La distribución de probabilidad necesita ser normalizada (las probabilidades deben sumar 100%). Ajustar los datos a una linea difícilmente va a ser identical a una distribución valida!
- Muchas funciones parecen lineales cued se grafican en escala log-log
- Discrepancias en la cola importan más que en cualquier otro lugar en un ajuste propuesto
- El "argumento de la estimación visual" no es una ciencia sólida



NETWORK SCIENCE

YOU'RE DOING IT WRONG

DIY.DESPAIR.COM

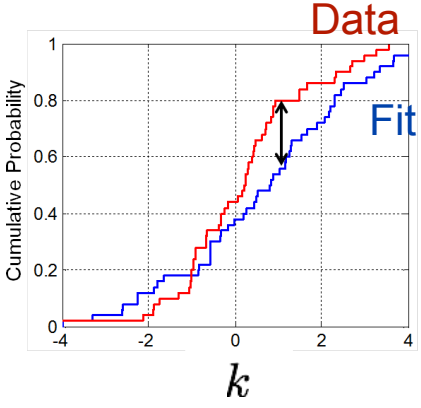


Clauset *et al.,* 2009

Ajuste de Power-law

La forma correcta: basado en máxima verosimilitud

Como encontrar K_{min}



Calcula la distribución acumulada de la da distribución de la data y de la data ajustada

Calcula la máxima distancia entre las dos distribuciones acumuladas

$$D = \max_{k \geq K_{min}} |S_k - P_k|$$

A.-L. Barabási. (2016). *Network Science, Chapter 4*.

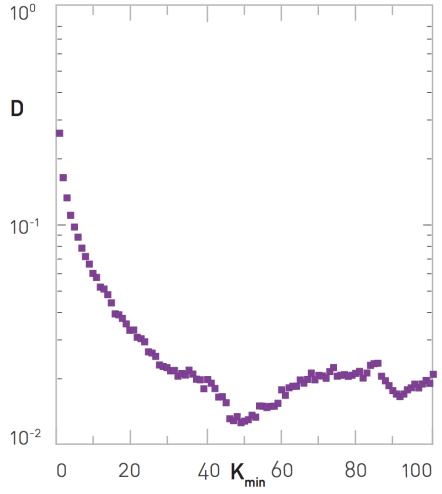
Ajuste de Power-law

La forma correcta: basado en máxima verosimilitud

k

Toma el K_{min} que minimiza

D



Código: http://tuvalu.santafe.edu/~aaronc/powerlaws/

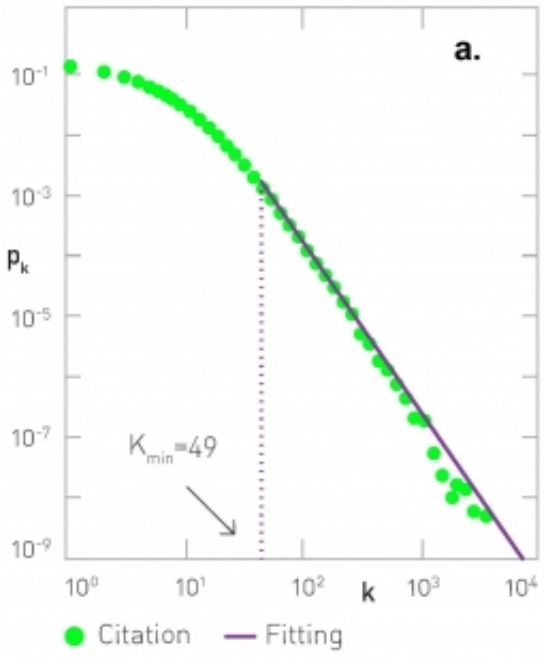
Scale-Free Networks Well Done

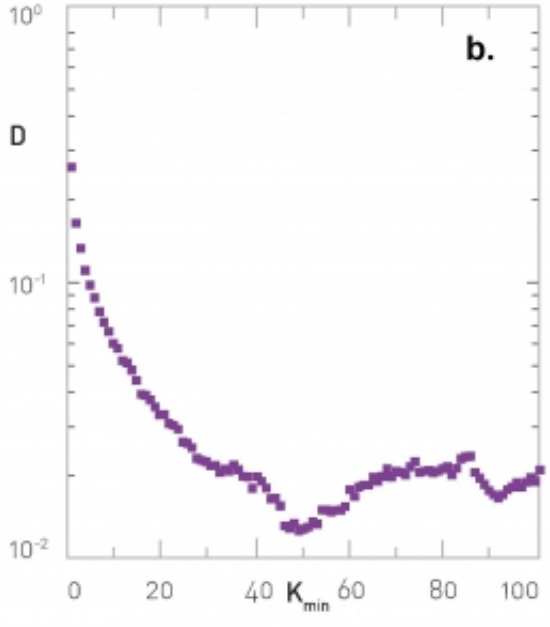
Ivan Voitalov,[1,2] Pim van der Hoorn,[1,2] Remco van der Hofstad,[3] and Dmitri Krioukov[1,2,4,5]

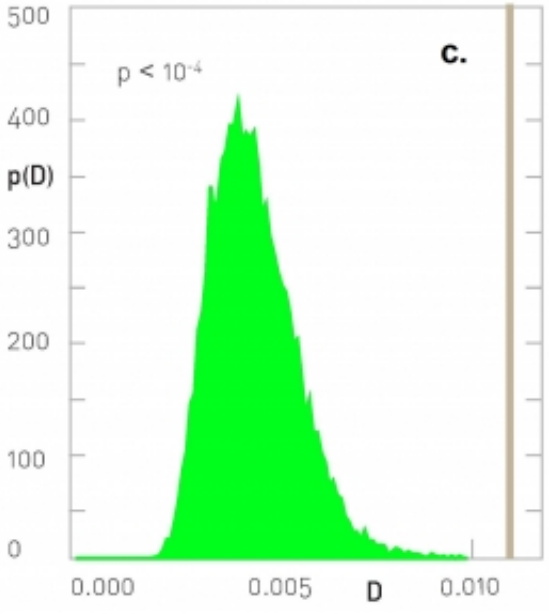
[1] *Department of Physics, Northeastern University, Boston, Massachusetts 02115, USA*
[2] *Network Science Institute, Northeastern University, Boston, Massachusetts 02115, USA*
[3] *Department of Mathematics and Computer Science, Eindhoven University of Technology, Postbus 513, 5600 MB Eindhoven, Netherlands*
[4] *Department of Mathematics, Northeastern University, Boston, Massachusetts 02115, USA*
[5] *Department of Electrical & Computer Engineering, Northeastern University, Boston, Massachusetts 02115, USA*

We bring rigor to the vibrant activity of detecting power laws in empirical degree distributions in real-world networks. We first provide a rigorous definition of power-law distributions, equivalent to the definition of regularly varying distributions that are widely used in statistics and other fields. This definition allows the distribution to deviate from a pure power law arbitrarily but without affecting the power-law tail exponent. We then identify three estimators of these exponents that are proven to be statistically consistent—that is, converging to the true value of the exponent for any regularly varying distribution—and that satisfy some additional niceness requirements. In contrast to estimators that are currently popular in network science, the estimators considered here are based on fundamental results in extreme value theory, and so are the proofs of their consistency. Finally, we apply these estimators to a representative collection of synthetic and real-world data. According to their estimates, real-world scale-free networks are definitely not as rare as one would conclude based on the popular but unrealistic assumption that real-world data comes from power laws of pristine purity, void of noise and deviations.

Procedimiento de ajuste





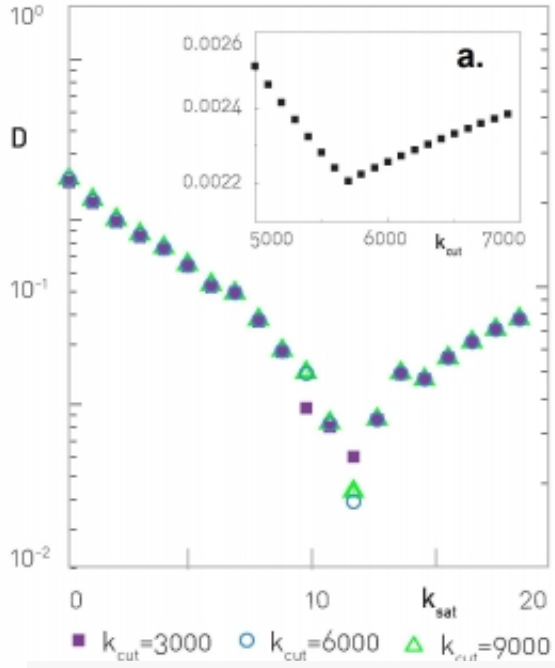


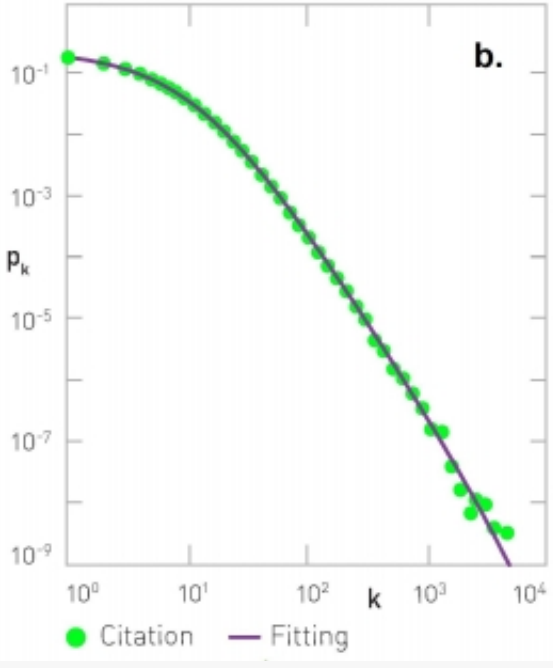
$$p_k = \frac{1}{\zeta(\gamma, K_{\min})} k^{-\gamma}$$

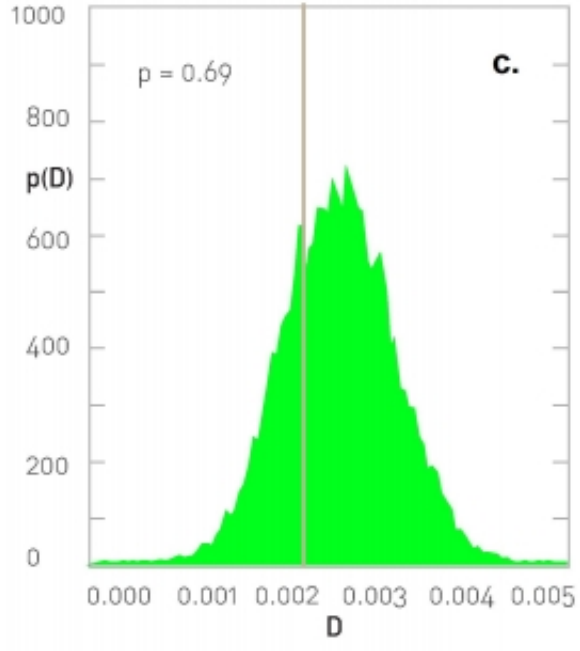
$$P_k = 1 - \frac{\zeta(\gamma, k)}{\zeta(\gamma, K_{\min})}$$

$$D = \max_{k \geq K_{\min}} |S(k) - P_k|$$

Procedimiento de ajuste



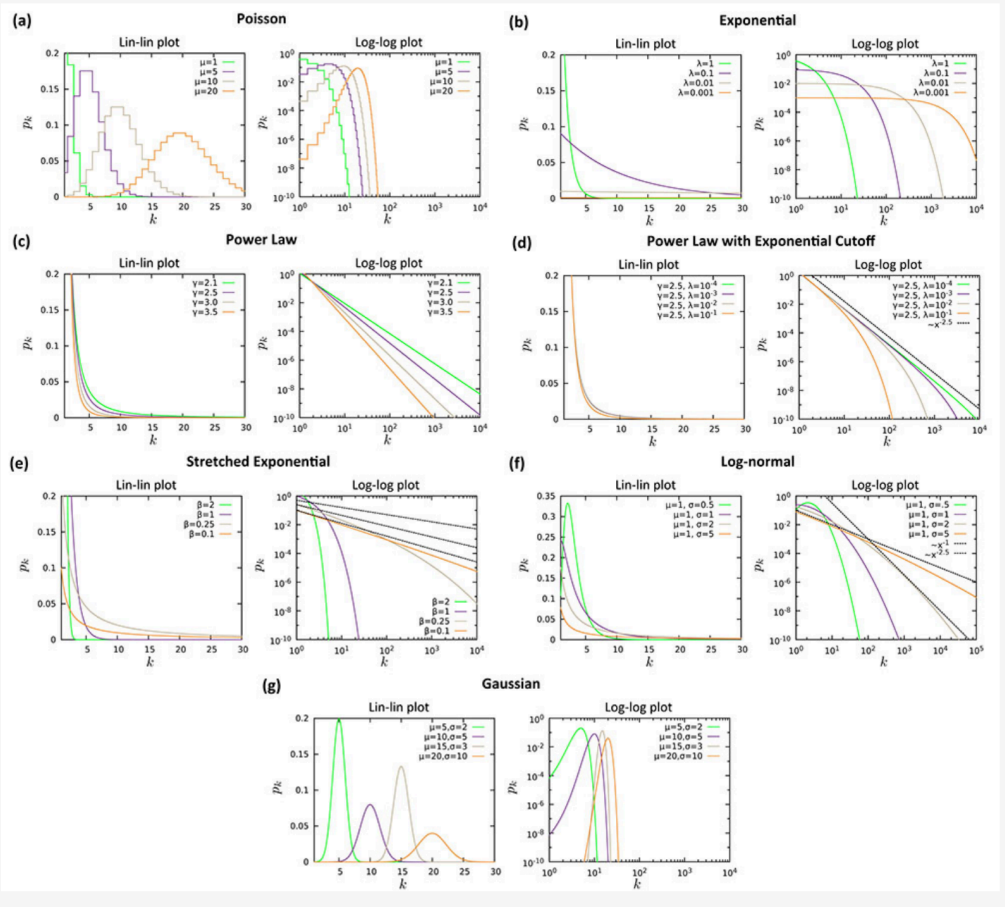




$$p_k = \frac{1}{\sum_{k'=1} (k'+k_{sat})^{-\gamma} e^{-k'/k_{cut}}} (k + k_{sat})^{-\gamma} e^{-k/k_{cut}} \qquad (4.47)$$

$$\log L(\gamma | k_{sat}, k_{cut}) = \sum_{i=1}^{N} \log p(k_i | \gamma, k_{sat}, k_{cut})$$

NAME	$p_x/p(x)$	$\langle x \rangle$	$\langle x^2 \rangle$
Poisson (discrete)	$e^{-\mu}\mu^x/x!$	μ	$\mu(1+\mu)$
Exponential (discrete)	$(1-e^{-\lambda})e^{-\lambda x}$	$1/(e^{\lambda}-1)$	$(e^{\lambda}+1)/(e^{\lambda}-1)^2$
Exponential (continuous)	$\lambda e^{-\lambda x}$	$1/\lambda$	$2/\lambda^2$
Power law (discrete)	$x^{-\alpha}/\zeta(\alpha)$	$\begin{cases} \zeta(\alpha-2)/\zeta(\alpha), & \text{if}\alpha > 2 \\ \infty, & \text{if}\alpha \leq 1 \end{cases}$	$\begin{cases} \zeta(\alpha-1)/\zeta(\alpha), & \text{if}\alpha > 1 \\ \infty, & \text{if}\alpha \leq 2 \end{cases}$
Power law (continuous)	$\alpha x^{-\alpha}$	$\begin{cases} \alpha/(\alpha-1), & \text{if}\alpha > 2 \\ \infty, & \text{if}\alpha \leq 1 \end{cases}$	$\begin{cases} \alpha/(\alpha-2), & \text{if}\alpha > 1 \\ \infty, & \text{if}\alpha \leq 2 \end{cases}$
Power law with cutoff (continuous)	$\frac{\lambda^{1-\alpha}}{\Gamma(1-\alpha)}x^{-\alpha}e^{-\lambda x}$	$\lambda^{-1}\frac{\Gamma(2-\alpha)}{\Gamma(1-\alpha)}$	$\lambda^{-2}\frac{\Gamma(3-\alpha)}{\Gamma(1-\alpha)}$
Stretched exponential (continuous)	$\beta\lambda^{\beta}x^{\beta-1}e^{-(\lambda x)^{\beta}}$	$\lambda^{-1}\Gamma(1+\beta^{-1})$	$\lambda^{-2}\Gamma(1+2\beta^{-1})$
Log-normal (continuous)	$\frac{1}{x\sqrt{2\pi\sigma^2}}e^{-(\ln x-\mu)^2/(2\sigma^2)}$	$e^{\mu+\sigma^2/2}$	$e^{2(\mu+\sigma^2)}$
Normal (continuous)	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/(2\sigma^2)}$	μ	$\mu^2+\sigma^2$



Procedimiento de ajuste

$k^{-\gamma}; [k_{min}, \infty]$					$(k+k_{sat})^{-\gamma}e^{-k/k_{cut}}$			
	γ	K_{min}	P-VALUE	PERCENT	γ	k_{set}	k_{cut}	P-VALUE
Internet	3.42	72	0.13	0.6%	3.55	8	8500	0.00
WWW (IN)	2.00	1	0.00	100%	1.97	0	660	0.00
WWW (OUT)	2.31	7	0.00	15%	2.82	8	8500	0.00
Power Grid	4.00	5	0.00	12%	8.56	19	14	0.00
Mobile Phone Calls (in)	4.69	9	0.34	2.6%	6.95	15	10	0.00
Mobile Phone Calls (out)	5.01	11	0.77	1.7%	7.23	15	10	0.00
Email-Pre (in)	3.43	88	0.11	0.2%	2.27	0	8500	0.00
Email-Pre (out)	2.03	3	0.00	1.2%	2.55	0	8500	0.00
Science Collaboration	3.35	25	0.0001	5.4%	1.50	17	12	0.00
Actor Network	2.12	54	0.00	33%	-	-	-	0.00
Citation Network (in)	2.79	51	0.00	3.0%	3.03	12	5691	0.69
Citation Network (out)	4.00	19	0.00	14%	-0.16	5	10	0.00
E.Coli Metabolism (in)	2.43	3	0.00	57%	3.85	19	12	0.00
E.Coli Metabolism (out)	2.90	5	0.00	34%	2.56	15	10	0.00
Yeast Protein Interactions	2.89	7	0.67	8.3%	2.95	2	90	0.52

p> 0,01 se considera estadísticamente significativo.

Estudiar el paquete powerlaws:
https://aaronclauset.github.io/powerlaws/